基于改进 BP 神经网络的 A 股投资实证研究

闫昱，李皓辰，通讯作者：王燕飞，经玲
（中国农业大学 理学院，北京 100083）

摘要

量化投资一直是国内外学者关注的热点问题，针对使用 BP 神经网络预测股票市场的相关问题，本文在对数据进行统计学上的数据预处理后，交替采用了遗传算法和模拟退火算法优化了 BP 神经网络的参数,并对 A 股市场中的股票进行预测。在此基础上，进行了为期十年的模拟交易，最终获得了 24.1%的年化收益率。

关键词：神经网络，优化算法，预测，投资
中图分类号：F830.91 文献标识码：A

1 引言

随着我国资本市场的不断完善，股票市场在国民经济中起着越来越重要的作用，投资于股票市场也成为了个人理财方式中最重要的方式之一，但是由于中国股票市场有一定的不健全方面，存在着诸多隐患，因此如果能找到一种合理的方式对股票的涨跌进行预测，指导人们进行合理的投资，就可以更好地促进我国资本市场进一步完善，推动我国国民经济朝着健康的方向发展。由此，寻找一种合理的预测方法有着极其重要的理论和实际意义。

2 项目背景

传统的股票市场中常用两种分析方法：技术分析法和基本面分析法，技术分析法是依据以往的交易相关指标例如 VOL,WACD 等对股票价格进行预测；基本面分析法是结合宏观政治经济条件和公司具体财务情况对股票的价值进行估计。但是技术分析法往往只适用于极短时间的预测，并且具有很强的不确定性；而基本面预测往往只能对中长期的涨跌情况进行揣测，并且对预测的时间段也很难有准确的估计。

另一方面，如何科学地使用数学模型对股票进行预测一直备受学者们的关注，国内外的学者也从不同的角度做出了许多方案对该问题进行探讨。针对股票市场的性质，影响股票价格波动的最重要的三类因素分别为股票的市场中的市场内部结构、经济基本面和政府政策。但是在具体的分析过程中，这三类影响因素在实际分析中有的可以具象为一些指标有些却难以定量研究。再加上其他形形色色的影响因素，使得股票的预测变得十分艰难。另外股票价格波动的非线性和股票信息的庞大信息量使得股票预测的算法复杂程度变得很高，而传统股市分析的技术分析法和基本面分析法在处理数个交易日内的走势预测时，难以 得到很好的预测效果。

由于股票具有非线性的特点，加之我国股票形势的特殊性，传统的统计学模型在 A 股市场的表现并不乐观。在参考文献[2]中，田利辉等人研究了五因子模型在中国证券市场中的表现，并且发现我国的定价因素与美国市场经验有较大区别。在参考文献[1]中，赵胜民等人研究了 Fama，French 在 2015 提出的五因子模型在中国 A 股市场中的表现，得出了虽然具有一定的有效性但是表现却不如三因子模型的结论。这些都是使用线性回归的方式对股票进行预测的研究。

除了经典的统计学方法，由股票是随着时间变化的这一性质，也有很多学者通过时间

序列的方式对股票进行预测，也有许多学者使用基于非统计学原理的创新算法对股票进行预测，比较具有代表性的有 GM、ANN 和 SVM 等等[3-6]，但是这些算法都有其局限之处，普遍具有的问题主要有易收敛到局部最优解、收敛速度慢、鲁棒性差和预测精度不足不足等等，所以很多学者提出了将两种以上的算法结合在一起的方案来克服单一算法的缺陷。例如有学者使用模拟退火算法对支持向量机的参数进行优化，也有学者使用布谷鸟算法对 BP 神经网络的参数进行优化[7]。另一方面国外也有很多学者进行了类似的研究[13-15]。

3 综述

本文将先通过经验选出十多个指标作为备选指标，为了使结论根据有一般性，即让模型能够更好的适用于整个 A 股市场，所以没有使用 PCA 等基于特征表示的降维方法，而是直接使用了特征选择即直接选出了几个指标作为使用的指标。在人工神经网络中，若是某一指标输入量级过大会导致整个算法收敛速度以指数级别变慢，并且由于我们所选择的指标之间量级差异巨大，所以先对数据进行了归一化处理。

考虑到股票市场复杂的非线性性质所以我们选择 BP 神经网络作为基础模型进行改进。

多层前馈式神经网络是目前应用比较广泛的神经网络，而 BP 算法是最著名的多层前馈网络训练算法。尽管随着神经网络科学的发展产生了许多优秀的算法，而 BP 算法本身又存在收敛速度慢易陷入局部极小值和推广能力差等不足，但由于其简单易行计算量小并行性强等优点，目前仍是多层前馈式网络训练的首选算法之一。并且已被人们广泛地应用于各种实际问题。以下把用 BP 算法作为网络学习算法的多层前馈式神经网络简称为 BP 网络。

BP 神经网络需要运用已有的数据进行训练，需要输入的数据有输入层、输出层和网络本身的参数。输出层比较简单，对于我们来说只需要预测之后一段时间的股票价格，而输入层我们选择是的是各项指标的预测值。我们的训练集就采用我们之前生成的关于股票价格的数据。

BP 神经网络的优势在于并不需要知道输入层与输出层具体的映射关系就可以通过对训练集的反复运算使用最速下降法，通过迭代法反复调整网络进而达到对训练集输出层的尽可能的拟合。

我们对所有的股票进行预测时采用的是同一个初始网络结构，这是为了让最终找到的网络结构更具有一般性。而另一方面为了提升模型的鲁棒性，在寻找最优的网络结构时，随机选择了二十只股票作为训练集，并以这二十支股票的预测精度为目标函数值，输入参数和网络结构中的阈值作为变量使用遗传算法和模拟退火算法进行交替优化进行寻优。

我们首先采用遗传算法求解全局最优解。由于优化算法层出不穷，优化理论也在不断完善、发展和应用中。在筛选优化算法的过程中，我们发现，各种优化方法可以分为启发式和随机式两大类。启发式算法如迭代法等具有较高的计算效率，但是容易陷入局部最优解；随机式算法如遗传等具有良好的全局寻优能力，但是计算时间较长。遗传算法以目标函数本身为适应度函数，无需求导、求逆等复杂运算，只需对变量进行编码处理，确定初始参数后即以与问题无关的方式求解问题，比较容易得到全局最优解和一批次优解，实现简单，并且能解决我们想图剖初始值局限这一因素，所以我们首先采取了遗传算法。我们在编程过程中也发现了其中遗传算法的不足，比如计算效率低，但由于我们这个求取最优解对时间的要求不是很高，所以我们认为用时间来换取更好的求全局最优的方案也是值得的。

分析数据结果，发现求解出来的局部最优解，之后考虑到遗传算法的局限性。它虽然解决了我们在非线性最优问题中遇到的初始值局限问题，但是常规遗传算法由于收敛速度慢、易早熟等缺陷，是有一定的概率进入局部最优解的。关键的是，一旦进入了局部最优解，就很难跳出了。实际上根本就跳不出，不管事先令其进化多少代，它们的后代都在那个局部范围内活动，从来不去新的空间探索。所以我们考虑再利用新的优化算法进行最优解求解本问题的最优解，我们采用了模拟退火算法。

优化之后再代入到整个市场的所有股票之中，若是效果良好，说明模型鲁棒性好，若是效果不佳则重新对二十只股票的网络结构进行优化，在进行了多次实验后我们终于得到了一个鲁棒性较好的模型。之后我们以此模型为基础，进行了为期十年的模拟交易，最终

发现在该模型指导下我们获得了 20%以上的收益。

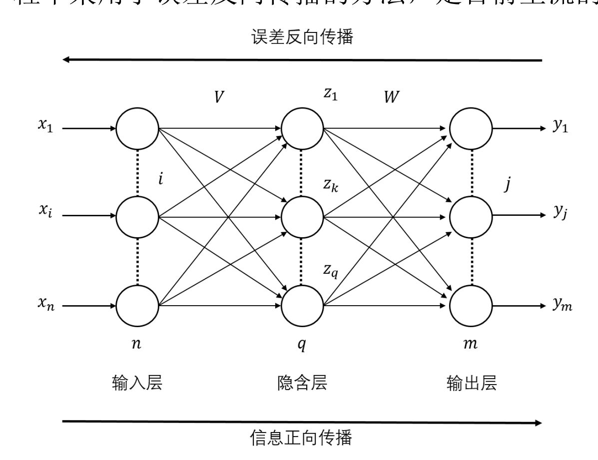
本文吸取了前人的研究经验，以 BP 神经网络对股票进行预测，并且使用了新型的优化算法对参数进行优化。从应用的目的出发，寻找了一个适用于所有股票的初始网络结构，以二十支股票作为训练组，以所有股票进行检验的方式充分说明了该模型具有很好的鲁棒性。

而在之前的研究中，大多对股票指数进行预测，有一部分文章对一部分股票进行了预测，但是往往模型是为了指导投资者进行投资，本文从实际出发，对上海交易市场中所有的股票都进行了实验，每二十个交易日进行一次买入和卖出的交易，并且考虑了所有的交易成本完全模拟实际投资，证明了我们所构建的模型在实际中的应用价值。

4 模型建立

2.1BP 神经网络简述

BP 神经网络是由 Rumelhart 等人在 20 世纪 80 年代提出的，其特点是在训练模型的过程中采用了误差反向传播的方法，是目前主流的神经网络模型。BP 神经网络的拓扑结构主要是包括一个输入层，一个或多个隐含层和一个输出层。每层中都包含着若干个神经元。BP 算法的学习的规则是最速下降法，学习过程由信号的正向传播和逆向传播两部分组成：在正向传播时，输入样本值先通过输入层传入各个隐含层，然后经过隐含层处理后，传向输出层。如果输出层的实际输出与期望的不一样，则将误差逆向传播，将误差反向传播给各层所有单元，从而得到各层单元的误差信号，并以此为依据修成各单元的权值，这种权值调整的过程是在不断进行的，就是网络的学习过程。这个过程持续到训练输出的误差减少到可以接受的程度终止，或者是运算次数达到或是预先设定的次数。



2.2 遗传算法的原理

遗传法则是根据适者生存原则选择下一代的个体。在选择时，以适应度为选择原则。适应度准则体现了适者生存，不适应者淘汰的自然法则。

1．初始化：选择一个群体，即选择一个串或个体的集合 bi，i=1，2，...n。这个初始的群体也就是问题假设解的集合。一般取 n＝30-160。通常以随机方法产生串或个体的集合 bi,i＝1，2，...n。问题的最优解将通过这些初始假设解进化而求出。

2．选择：根据适者生存原则选择下一代的个体。在选择时，以适应度为选择原则。适应度准则体现了适者生存，不适应者淘汰的自然法则。给出目标函数 f，则 f(bi)称为个体 bi 的适应度。以方程

为选中 bi 为下一代个体的次数。满足适应度较高的个体，繁殖下一代的数目较多；适应度较小的个体，繁殖下一代的数目较少，甚至被淘汰。

这样，就产生了对环境适应能力较强的后代。对于我们的模型来说，就是选择出和最优参数值较接近的中间解。

2.3 模拟退火算法的原理

模拟退火算法源于对固体退火过程的模拟，用一组称为冷却进度表的参数控制算法的进程，使算法在控制参数 t 徐徐“降温”并趋于零时，最终求得组合优化问题的相对全局最优解。其中优化问题的一个解 i 及其目标函数 f（i）分别与固体的一个微观状态 i 及其能

量 E_i 相对应。令随算法进程递减的控制参数 t 担当固体退火过程中温度 T 的角色，则对于 t 的每一个取值，算法采用 Metropolis 接受准则，持续进行“产生新解--判断--接受/舍弃”的迭代过程而达到该“温度”下的“平衡点”。由于搜索是随机的，会经常遇到较坏的点，但模拟退火算法并不像其他优化算法那样只接受好的点，它会按照 Blotzmann 分布规律不时的接受恶化解，从而跳出局部极小的“陷阱”。

2.3 利用遗传算法和模拟退火算法寻找输入参数和阈值

用遗传算法和模拟退火算法学习神经网络的权重，也就是用遗传算法和模拟退火算法来取代传统的学习算法。学习算法的评价标准是：简单性、可塑性和有效性。然而一般情况下，这前两者存在一定程度上的矛盾：简单的算法并不有效，可塑的算法又不简单；有效性的算法则要求算法的专一性、完美性，从而又与算法的可塑性、简单性相冲突。目前广泛研究的前馈网络中采用的是 Rumelhart 等人推广的误差反向传播(BP)算法，BP 算法具有简单和可塑的优点，但是 BP 算法是基于梯度的方法，这种方法的收敛速度慢，且常受局部极小点的困扰，我们采用了遗传算法和模拟退火算法来优化了神经网络的参数和阈值，实质上是把神经网络的权值学习和结构优化结合起来求解。可以显著地起到提高收敛速度和避免局部最优解的效果。同时，在计算中我们发现遗传算法和模拟退火算法之间也存在着优缺点互补的现象，于是我们采用了交替使用两种算法，力求在计算速度和计算精度上找到一个平衡点，来实现我们对股市快速有效的预测。

5 实证研究

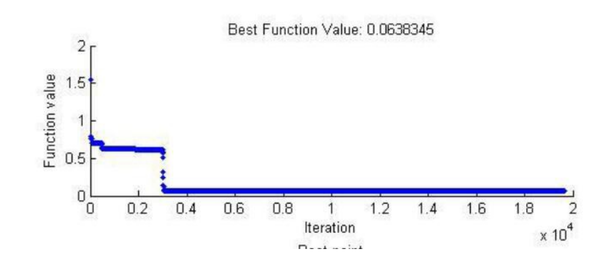
在本文之中，使用的所有数据均来自于 wind 软件（wind 是一款可以通过 matlab 对接的十分有效的金融资讯平台），股票样本主要选取 2006 年 6 月 27 日到 2016 年 6 月 27 日的时间序列数据，并且选取的是上海交易所上市的公司。我们选取了如下的十七个指标作为回归数据的参考，并且对相关系数做出了如下计算:

特征名称	相关系数
前收盘价	-0.0107063
开盘价	-0.0104103
最高价	0.0161716
最低价	0.0158218
收盘价	0.0428903
成交量	0.0897643
成交额	0.070033
涨跌幅	0.8150088
振幅	0.0650866
均价	0.0222136
换手率	0.1100169
持仓量	0
持仓量变化	0
相对发行涨跌	0.035824

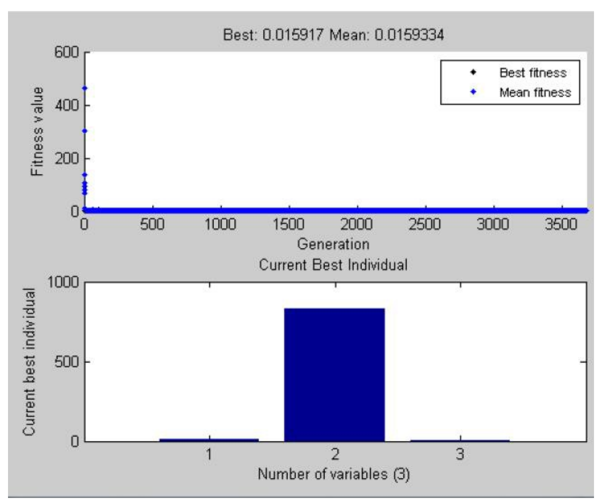
幅	6
净流入资金	0.3320283
市盈率	0
市净率	0

在计算之后，我们选取相关系数绝对值最大的十个指标作为备选指标，

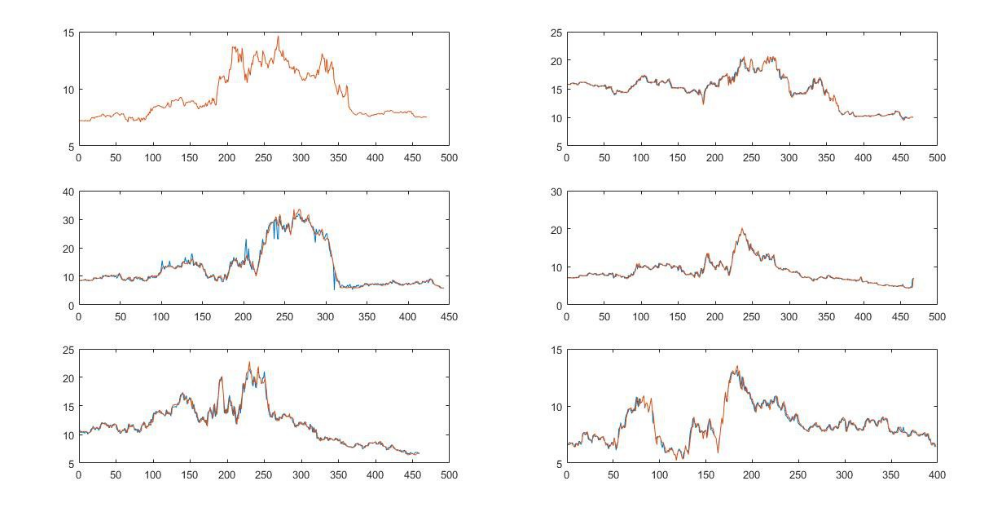
我们使用遗传算法的优化效果如下，该图表示我们进行了接近两万次的迭代，可以看出在 3000 次迭代之后就误差在已经不再下降，



之后我们再使用模拟退火算法对神经网络的参数进行优化。进行优化效果如下，可以看出只在一开始的时候模型预测的误差有所下降，之后的迭代并没有使误差得到有效的下降。

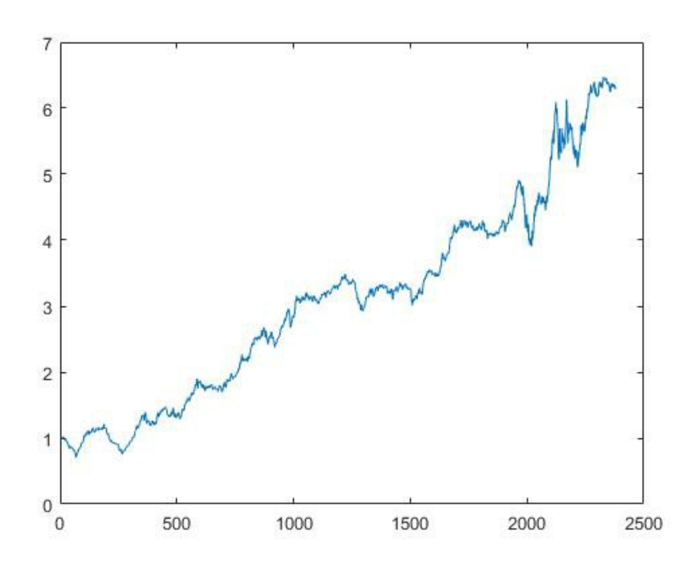


为了便于观测，我们选择中国石油、苏宁云商、大智慧、兴业银行、苏宁云商和蒙草抗旱的预测曲线在最近四百多天的价格的对比图，最终达到预测准确度 99%，即误差在 1% 左右。



下面给出中国石油的部分日期的部分指标的真实值与预测值的对比：

日期	开盘价	最高价	收盘价	交易量（股）	涨跌幅	预测收盘价
2016/5/27	7.21	7.22	7.2	9075211	-0.01	7.199132
2016/5/26	7.21	7.25	7.22	14301584	0.01	7.254899
2016/5/25	7.22	7.24	7.19	11116979	-0.03	7.186152
2016/5/24	7.19	7.22	7.22	14916983	0.03	7.236926
2016/5/23	7.21	7.23	7.2	10559126	-0.01	7.19459
2016/5/20	7.17	7.21	7.21	8715582	0.04	7.244347
2016/5/19	7.18	7.23	7.18	16951884	0	7.204209
2016/5/18	7.21	7.22	7.2	28808330	-0.01	7.223426
2016/5/17	7.21	7.25	7.24	17886829	0.03	7.243347
2016/5/16	7.17	7.2	7.2	13220329	0.03	7.230929
2016/5/13	7.2	7.25	7.19	21040632	-0.01	7.206081
2016/5/12	7.21	7.21	7.2	17221805	-0.01	7.19605
2016/5/11	7.21	7.25	7.22	14851902	0.01	7.235215
2016/5/10	7.18	7.22	7.19	13713153	0.01	7.202872
2016/5/9	7.32	7.32	7.2	34754111	-0.12	7.202128
2016/5/6	7.46	7.46	7.34	32726175	-0.12	7.332826
2016/5/5	7.45	7.48	7.46	17745527	0.01	7.449884
2016/5/4	7.48	7.51	7.46	21305853	-0.02	7.452454
2016/5/3	7.4	7.51	7.51	29245563	0.11	7.522083



这是从 2016 年 6 月 27 日往前 2380 天同时持有 100 值股票的带有对冲的回测数据，从收益上来看获得了接近百分之 24%的年化，夏普比率达到了 1.1086。并且这一次回测是以散户为基准的，所以每天都有千分之一的交易成本，如果是机构投资者这一成本将会大大降低可以获得更高的收益。

在交易过程中，在开始的 2000 天左右的时候曲线具有较大的起伏，这正好与 2015 年下半年的震荡相吻合，而 1200 天左右的波动正好与股票市场的低迷时期相吻合，所以虽然采用了对冲的策略，但是在具体的投资过程中往往没法应对股灾是情境。

为了避免实验的偶然性，我们还对四组数据每组数据试验了 60 支到 150 支股票，并且每一组收益都在百分之十五以上。

6 结论

股票的预测理论一直是学者们所关注的热点，从传统统计学、经济学和机器学习等学科都可以提出不同的预测理论，本文所探讨的就是利用一种优化神经网络进行预测的准确性和利用此模型进行投资的有效性。

在本文之中，首先我们考虑了十七个常用指标的有效性，并且通过十年的数据得到了较为合理的结论，这一结论可以为很多预测理论提供一定的参考。之后我们建立了一个反向传播误差人工神经网络来对股票数据进行预测，但是神经网络的最难以攻克的问题就是其初始参数繁多难以给出，作者使用了两种优化算法对神经网络的参数进行交替迭代优化最终我们获得了拥有较高收益量化模型。为了能够更直观的体现预测精度，作者给出了六只股票预测值与真实值之间的对比图，但是神经网络也有一些其固有的弊端，首先就是其计算时间较其他模型较长，如果再考虑其参数的优化，那么需要训练成千上万次网络，所以需要很久的计算时间。

从应用角度出发，本文给出了一个十分有效的投资模型；该模型指导下的投资不仅收益可观，而且收益十分稳定，可以用于各类投资领域。同时佐证了中国证券市场具有弱有效性。不仅对现实之中的优质选股问题提供了一定的方案，同时该问题还可用到公司财务评估、水质污染计量和土地资源评估等各个领域。

参考文献

[1] 赵胜民,闫红蕾,张凯 Fama-French 五因子模型比三因子模型更胜一筹吗——来自中国 A 股市场的经验证据 [J].南开经济研究,2016（2）

[2] 田利辉,王冠英.我国股票定价五因素模型：交易量如何影响股票收益率？[J].南开经济研

究,2014（2）
[3] 吴玉霞,温欣.基于 ARIMA 模型的短期股票价格预测 [J].统计与决策,2016（23）
[4] 黄宏运,王梅,朱家明.基于多元回归分析的多因子选股模型 [J].通化师范学院学报（自然科学）,2016（4）
[5] 张玉川,张作泉.支持向量机在股票价格预测中的应用 [J].北京交通大学学报,2007（6）
[6] 王维贤,陈利军.股票价格预测的建模与仿真研究[J].计算机仿真,2012（1）
[7] 孙晨,李阳,李晓戈,于娇艳.支基于布谷鸟算法优化 BP 神经网络模型的股价预测 [J].计算机应用与软件,2016,33（2）
[8] 吴微,陈维强,刘波.用 BP 神经网络预测股票市场涨跌 [J].大连理工大学学报,2001 41（1）
[9] 张玉川,张作泉.支持向量机在股票价格预测中的应用 [J].北京交通大学学报,2007（6）
[10] 智晶,张冬梅,姜鹏飞.基于主成分的遗传神经网络股票指数预测研究[J].计算机工程与应用,2009 56（3）
[11] 贺本岚.股票价格预测的最优选择模型[J].统计与决策,2008,(6).
[12] 游士兵,都娟.胜算指标在股票技术分析中的有效性检验[J].统计与决策,2016,(22)
[13] ERUMELHART D.E,HINTON G.E, WEILLIAMS R,J.Learning internal representations by error-propagation [A].Parallel Distributed Processing:Vol ,[M].Cambridge MA:MIT Press,1986.
[14] Fama E.F.,French K.R.A Five-Factor Asset Pricing Model [J]. 通 Journal of Financial Economics,2015,116（1）
[15] J.Patel， S.Shah， P.Thakkar and K.Kotecha.Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques [J]. Expert Systems with Applications. 2015, 42(1):259–268.